Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

Why Conspiracy Theory Communities Endure: The Interplay of Feedback, Habits, and Community Context

Veronika Batzdorfer,[1] **Mattia Samory,**[2] **Sven Banisch**[1]

[1] Karlsruhe Institute of Technology
[2] Sapienza University of Rome
veronika.batzdorfer@kit.edu, mattia.samory@uniroma1.it, sven.banisch@kit.edu

Abstract

Online conspiracy theory (CT) communities pose challenges for digital governance, yet the behavioral mechanisms sustaining their participation remain unclear. We analyze 2.1 million posts from 2,711 users across 4,435 subverses on Voat, an alternative platform with minimal moderation.

Using survival analysis and hierarchical mixed-effects models, we show that habitual posting reduces sensitivity to social feedback, decoupling participation from the upvotes and downvotes that users receive. The type of content that users contribute varies in its effect: while group-oriented CT signals attenuate community engagement, trait-oriented signals, such as allegations of secrecy, have little effect. While posting habit is associated with disengagement from subverses in general, ideological alignment mitigates this effect in CT subverses. Together, these results reveal a multi-scale framework in which social reinforcement, routine participation, and community context interact to sustain CT engagement. By isolating mechanisms beyond CT alignment, we explain how fringe communities persist despite volatile rewards, providing new insights into the structural resilience of CT spaces.

Introduction

Conspiracy theories (CTs) are widespread in public discourse and have tangible real-world consequences, including their capacity to mobilize individuals and communities (Douglas and Sutton 2022). CTs typically manifest as complex narratives involving multiple actors, presumed motives and hidden intentions, perceived threats, and efforts to conceal causality (Phadke, Samory, and Mitra 2021a). Recent episodes, such as the role of online CT groups in the U.S. Capitol attack, have drawn attention to how digital platforms can facilitate ideological radicalization and offline mobilization. At the same time, debates surrounding content moderation policies and platform governance have intensified concerns about how online communities may cultivate sustained engagement with CT content (Papakyriakopoulos, Serrano, and Hegelich 2020).

A growing body of research has examined various facets of online CT communities, characterizing the variety of CT content they promote and distinguishing it from both non-CT and debunking rhetoric (Diab, Nefriana, and Lin 2024; Korenčić et al. 2024; Samory and Mitra 2018). Further research explored the social and interactive aspects of CT, such as the factors that influence users' entry into CT communities, the longitudinal evolution of user engagement within them, as well as the broader interaction patterns between CT and non-CT communities (Phadke, Samory, and Mitra 2021b, 2022; Russo, Ribeiro, and West 2024). Together, these studies offer a foundational understanding of user behavior, content production, and community dynamics in CT contexts. However, the mechanisms linking content exposure, social reinforcement, and habitual participation remain poorly understood (Anderson and Wood 2023; Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014).

Here, we provide an empirical and theoretical framework for multi-level engagement in CT communities, integrating psychological theories of habit formation and reinforcement learning. Habit theory predicts that repeated behaviors become increasingly automatic, reducing sensitivity to immediate rewards (Bermúdez and Felletti 2021). Reinforcement learning posits that behavior adapts to the value of social feedback, yet community cohesion can buffer or amplify these effects (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014). Unlike prior work primarily situated in mainstream platforms like Reddit and Twitter, we study Voat, an alternative social media platform. Voat's distinct moderation policies and community norms provide a unique opportunity to study user behavior in a less-curated environment, particularly within ideologically cohesive CT communities. In contrast to past work, we study how engagement mechanisms differ between CT-aligned and non-CT spaces.

Our central contribution is to frame **engagement as a multiscale process** that unfolds across timescales (Fig. 1): from immediate responsiveness at an hourly rhythm, through weekly shifts in activity, to monthly persistence or dropout. We conceptualize these dynamics as seesaws between goal-directed expectations and habitual tendencies. At the micro-level, responsiveness reflects the balance between sensitivity to feedback and the automaticity of habitual responses. At the meso-level, engagement is structured by exploratory shifts into community spaces, guided less by immediate reinforcement than by repeated exposure and narrative group cues. At the macro-level, persistence and dropout capture durable commitments sustained by entrenched routines and CT communities.

With this framework, we show three key insights, depicted in Figure 1B:

- **Micro-level feedback attenuation**: Users with stronger posting habits show reduced responsiveness to social reinforcement, reflecting a shift from reward-driven to routine-driven engagement.
- **Meso-level selective upshifts**: Transitions into increasingly CT spaces are structured less by immediate feedback than by habitual exposure and narrative cues.
- **Macro-level conditional persistence**: While intense posting routines can undermine long-term persistence, alignment within CT communities stabilizes engagement.

Our findings highlight how stimulus–response associations and goal-directed expectations jointly shape engagement, with habits serving as the bridge between immediate posting decisions and longer-term immersion in CT communities. This perspective resonates with related work showing that repeated participation can intensify warning behaviors of radicalization, while community context and feedback dynamics determine whether such behaviors are reinforced or diminished (Habib, Srinivasan, and Nithyanand 2022). We provide the code[1] for reproducibility.

Related Work

Rewards and Online Engagement. Diverse motivational factors drive engagement on online platforms, including social drivers such as homophily and group identity. From a classical learning theory perspective, platform features can function as secondary rewards, serving as proxies for primary social rewards like belonging and recognition (e.g., upvotes, re-sharing, received comments) (Turner et al. 2025). Prior research has adopted a reinforcement learning framework to examine the temporal dynamics of posting behaviors on Instagram, showing that time intervals between posts correlate with the magnitude of previously received rewards (Lindström et al. 2021). However, user motivations are not static; they evolve, particularly through repeated behaviors such as posting, commenting, or swiping (Turner et al. 2025; Grinberg et al. 2016). As these actions are repeated, they may become more automatic, driven by contextual cues (e.g., ephemeral comment threads, notifications) rather than deliberate cognitive processes (Anderson and Wood 2023; Turner et al. 2025). This automaticity diminishes the user's sensitivity to rewards, reflecting a shift in underlying motivations (Anderson and Wood 2023). Our work builds on these insights by explicitly modeling feedback responsiveness as a function of habit formation, linking moment-to-moment reward sensitivity with longer-term migration and persistence in CT-aligned spaces (Fig. 1).

Habit and Routinization. Habitualized behaviors may also play a critical role in engagement with conspiracy-related content, as suggested by previous research on moral outrage (Brady et al. 2021). Brady, McLoughlin, Doan, and Crockett (2021) demonstrated that, on Twitter and in experimental settings, positive feedback on moral outrage expressions predicts future outrage and that prevailing platform contexts shape users' sensitivity to social feedback in the form of norm learning. Our approach situates habit formation alongside reinforcement learning and community influences, allowing us to capture how routines modulate responsiveness to social feedback, drive exploration across communities, and stabilize long-term engagement (see Fig.1 A).

Conspiracy Beliefs. Conspiracy narratives can be interpreted as manifestations of evolved cognitive mechanisms that historically served to detect hidden intentions, anticipate threats, and infer agency in social environments (Van Prooijen and Van Vugt 2018). In contemporary digital platforms, these mechanisms may become maladaptive, producing overgeneralizations, heightened threat perception, and illusory pattern detection (Van Prooijen and Van Vugt 2018). We hypothesize such cognitive biases to interact with social contexts to shape how individuals participate in conspiratorial discourse.

We operationalize these mechanisms through narrative content features that capture both cognitive and social dimensions of conspiratorial messages. **Trait cues**, including *secrecy*, *pattern*, and *threat*, reflect intrinsic properties of the message itself, signaling hidden influence, repeated orchestrated patterns (Whitson and Galinsky 2008), or imminent danger. These cues may instigate epistemic uncertainty, motivating extended interpretive search and engagement, while also reflecting cognitive dispositions such as illusory pattern perception (Whitson and Galinsky 2008), agency detection, and paranoid reasoning (Van Prooijen and Jostmann 2013). They may further capture elements of self-disclosure, revealing the poster's internal beliefs, expectations, and management of uncertainty through conspiracy narratives (Van Prooijen and Jostmann 2013).

In contrast, **group cues**, encompassing *actors* and *actions*, focus on agent-centered cues that identify coalitions, assign responsibility, and describe coordinated behaviors (Van Prooijen and Jostmann 2013). By highlighting actors framed as malicious or instrumental in harmful activity, these features may provide attributional closure and anchor users within existing community narratives. Together, trait and group cues articulate complementary paths. At the community level, these content signals may aggregate to exploration, immersion, and persistence patterns. By linking post-level cognitive and social signals to engagement upshifts and longitudinal engagement patterns (Waller and Anderson 2021; Phadke, Samory, and Mitra 2021b), we trace how individual cognitive mechanisms scale into action selection.

Data

Voat was an online platform modeled after Reddit that allowed users to interact in a variety of thematic communities, called subverses (Mekacher and Papasavva 2022). Discussions were structured as trees, rooted in a post called submission, to which replying posts, called comments, could develop into conversation threads. Voat presented itself as a "free speech alternative" platform to Reddit, in opposition

[1] https://github.com/nika-akin/ICWSM-CT-multilevel

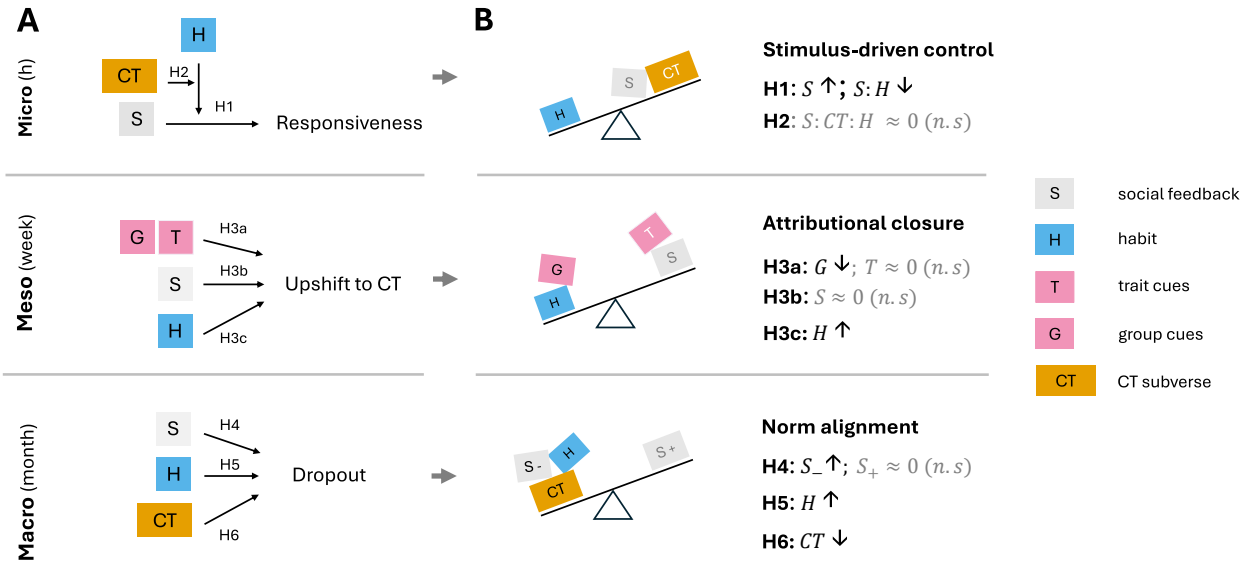

Figure 1: Multiscale mechanisms of engagement with online CT communities. **(A)** Conceptual framework across three timescales: responsiveness to social feedback (micro, hourly time scale), upshift to CT subverses (meso, weekly), and subverse dropout (macro, monthly). **(B)** Empirical results and dual-system interpretation. Seesaws lean toward stronger influences, versus non-significant and less central effects. Top: As posting habits consolidate, users react less to social rewards, regardless of whether in CT communities. Middle: Transition toward CT communities is shaped less by trait cues in users' content (allegations of secrecy, identification of patterns, claims of threats) than by group cues (explication of actors and actions in post narratives). Bottom: Although generally, negative social feedback increases the likelihood of exiting a community and habitual posting accelerates the process, ideological alignment with CT communities sustains engagement.

to the latter's norms and moderation practices, which were perceived as excessively restrictive.

We drew a sample of $N = 2,711$ users following the methodology in prior work (Mekacher and Papasavva 2022). We restricted sampling to unique posts, with at least 120 characters and without external links, to identify substantial on-platform contributions. From these, we drew a sample of users, stratified across seed subverses, which were identified in previous literature and span both CT (anon, conspiracy, pizzagate, greatawakening, theredpill) and non-CT topics (news, gaming, science, showerthoughts) (Mekacher and Papasavva 2022; Phadke, Samory, and Mitra 2021b).

Ethics. We restricted analyses to users with persistent registered usernames, excluding anonymous identifiers that change with each post, to enable longitudinal study of participation over time. Usernames were treated as random identifiers only; no information in the dataset can link online accounts to offline individuals. All analyses were conducted in aggregate, and no redistribution of raw data, usernames, or identifiable text excerpts occurs. To further minimize risks of re-identification, we report only community-level results. Our approach follows established ethical guidance for research using online data (Fiesler and Proferes 2018; Proferes et al. 2021; Bruckman 2002).

RQ1: How Does Responsiveness to Social Feedback Vary with Routine Behavior and Community Context?

Our starting point is the **micro-level**: how quickly users respond to signals from peers in their immediate environment. **RQ1** asks whether the timing of users' posting behavior—operationalized as responsiveness to feedback (i.e., posting latency)—varies with habit and community context. This focus on responsiveness highlights the most short-term dynamics of participation, where reinforcement signals (e.g., upvotes, downvotes) can nudge the tempo of engagement. A central principle in behavioral psychology and reinforcement learning is that behavior is shaped by the perceived value of outcomes: higher expected rewards accelerate action, whereas lower rewards slow it down (Herrnstein 1970). On online platforms, upvotes or replies operate as such rewards, often encouraging faster re-engagement (Lindström et al. 2021; Banisch et al. 2025). Habit theory predicts that as behaviors become routinized, their sensitivity to rewards diminishes (Bermúdez and Felletti 2021). Once participation becomes habitual, posting may occur largely on "autopilot," guided by internalized norms rather than immediate reinforcement. Contextualizing this in CT spaces, the rationale is that CT communities might cultivate stronger community expectations, which buffer against short-term reinforcement signals and instead sustain participation through shared identity and ideology. These concepts inform our hypotheses:

- **H1 (Routine-response):** As users' posting habits H strengthen, the effect of social feedback S on posting latency diminishes.
- **H2 (Subverse-response):** The routine-driven reduction in feedback sensitivity (S) may differ depending on the type of subverse (SV_{type}). Users in CT-aligned subverses are hypothesized to show weaker responsiveness to feedback compared with users in non-CT subverses.

Measures

Outcome: Posting Latency. Responsiveness to feedback was measured as the time between consecutive posts by the same user (in hours), log-transformed to reduce skewness. Because habits are more likely to be enacted under short response-preparation times, when goal-directed control has less leverage (Buabang et al. 2025), posting latency provides a behavioral proxy for the balance between routinized and deliberative action. In this operationalization, users who return quickly and repeatedly are more likely to act from habit rather than deliberate choice. Main predictors were *posting routines* and *lagged net votes*. Controls included *account tenure*, *subverse growth*, and *subverse type* (CT vs. non-CT).

Habit Formation

Posting Routines quantify how routinely and consistently users engaged with the subverse. We identify *occasional* users as those below the 75$^{\text{th}}$ percentile of daily posting frequency.[2] Among the remaining regular contributors, we distinguish between *steady* and *bursty* users based on their posting pattern. We calculated a burstiness coefficient B using the inter-post gap, i.e., the time difference in hours between consecutive posts:

$$B = \frac{\sigma_{\text{gap}} - \mu_{\text{gap}}}{\sigma_{\text{gap}} + \mu_{\text{gap}}}$$

where μ_{gap} and σ_{gap} denote the mean and standard deviation of inter-post gaps, respectively. We then categorized regular users' habits as *bursty* if $B > 0.7$,[3] *steady* otherwise. (Alternative thresholds—50th, 66th, 75th, and 90th percentiles of posts-per-day—yielded qualitatively identical results.)

Account Tenure. To account for life-cycle effects of persistence, we controlled for the number of days between user registration and platform shutdown, assigning users to one of three groups: *short* if tenure $\leq$ 365 days, *mid* if 365 $<$ tenure $\leq$ 1000 days, and *long* if tenure $>$ 1000 days.

Social Feedback

Net Votes. Peers can attribute social feedback to users' comments by up- or downvoting them. When considering latency to the next post, lagged net votes represent the difference between the number of upvotes and downvotes a post receives, reflecting overall peer reward.

[2]For simplicity, we do not distinguish between users' comments and submissions, and will refer to both as "posts."

[3]The 0.7 threshold was selected to balance bursty and steady users; we tested other thresholds, which led to comparable results.

Community Context

Subverse Type (CT vs. Non-CT). To effectively characterize different subverses according to their conspiracy theory (CT) dimensions, it is crucial to recognize that users may engage with multiple, potentially contradictory CTs over different communities (Phadke, Samory, and Mitra 2021b). The notion of overlapping engagement in multiple communities has been crucial to understanding the success of communities and discussions (TeBlunthuis et al. 2022). Our approach builds upon the methodology of co-membership outlined by Waller and Anderson (2021). We first filtered out users with insufficient posting frequency across various subverses and constructed a user-by-subverse matrix and embedded the communities with Word2Vecf (Levy and Goldberg 2014). To simplify the data and highlight commonalities based on user interactions, we performed Principal Component Analysis (PCA) and selected the first principal component, which accounted for the majority of the variance. Subsequently, we standardized these scores using z-transformation to ensure comparability. We then projected the normalized subverse vectors on the CT dimension vector (consisting of seed CT communities such as *v/conspiracy*) to assign each subverse a score along the CT dimension. For example, a subverse similar to the CT seed will have a higher score in this continuum. For any value of the embedding score greater than 0, we considered the subverse more conspiracy-affiliated, else non-conspiracy. (For a validation, see Appendix A (Fig. 5)).

Subverse growth. We measure the average growth rate of a subverse as the ratio of its number of subscribers to its age in days:

$$\text{growth rate} = \frac{\text{subscriber count}}{\text{sv age (days)}}$$

where *sv age (days)* is the number of days between the subverse's creation and the Voat shutdown date (December 25, 2020). Including subverse growth as a control accounts for differences in community size dynamics: faster-growing subverses may offer stronger reinforcement signals and greater visibility, which could independently affect user responsiveness and participation tempo. By adjusting for this factor, we ensure that the effects of habit class or subverse type are not confounded by the underlying growth trajectory of the community.

Model

To test whether routines moderate the effect of social feedback (H1) and whether this moderation varies by subverse type (H2), we fitted a linear mixed model with a three-way interaction:

$$\begin{aligned}\log(1+Y_{ijt}) = \alpha &+ \beta_1 S_{ijt} + \beta_2 H_i + \beta_3 SVtype_j \\ &+ \beta_4 S_{ijt} H_i + \beta_5 S_{ijt} SVtype_j \\ &+ \beta_6 H_i SVtype_j + \beta_7 S_{ijt} H_i SVtype_j \\ &+ \beta_8 G_{jt} + \beta_9 T_{it} + u_i + v_j + \varepsilon_{ijt},\end{aligned}$$

$$u_i \sim \mathcal{N}(0, \sigma_u^2), \quad v_j \sim \mathcal{N}(0, \sigma_v^2) \tag{1}$$

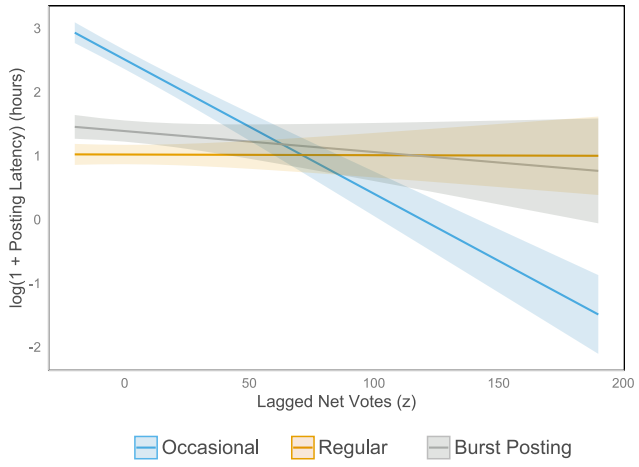

Figure 2: Predicted feedback responsiveness (log-transformed posting latency) by standardized lagged net votes across posting routines (habit classes). Shaded bands indicate 95% confidence intervals.

Where $\log(1 + Y_{ijt})$ is the log-transformed posting latency of user i in subverse j at time t, S_{ijt} is lagged net votes, H_i is habit class, $SVtype_j$ indicates subverse type (CT vs. non-CT), G_{jt} is subverse growth, , T_{it} is tenure, and u_i, v_j, ε_{ijt} are user, subverse, and residual random effects, respectively.

Results

Random intercepts for users (n = 2,586) and subverses (n = 4,149) captured individual and community-level heterogeneity. Variance was higher at the user level ($\hat{\sigma}^2_u = 0.52$, SD = 0.72) than at the subverse level ($\hat{\sigma}^2_v = 0.02$, SD = 0.13). Residual variance was $\hat{\sigma}^2 = 1.80$ (SD = 1.34). Conditional $R^2 = 0.354$; fixed effects alone accounted for $R^2_{\text{marginal}} = 0.162$.

H1: Routine-response. Supporting H1, lagged net votes predicted shorter posting latency ($\hat{\beta}_S = -0.02$, 95% CI [-0.02, -0.02], $p < .001$) though the effect size was relatively modest. In contrast, habitual users posted faster overall and regular posters ($\hat{\beta}_H = -1.49$, 95% CI [-1.57, -1.41], $p < .001$) more so than bursty ones ($\hat{\beta}_H = -1.12$, 95% CI [-1.21, -1.04], $p < .001$). Interactions between lagged net votes and habit indicated slightly reduced feedback sensitivity among regular posters ($\hat{\beta}_{S:H} = 0.021$, 95% CI [0.02, 0.02], $p < .001$) and bursty posters ($\hat{\beta}_{S:H} = 0.018$, 95% CI [0.02, 0.02], $p < .001$) (Fig. 2). This pattern (Fig. 2) mirrors prior work by Anderson and Wood (2023), who showed on Instagram that for users with weak posting habits, higher social rewards (likes and loves) are associated with shorter latencies between posts (negative slope), whereas for users with strong habitual posting, increased rewards lead to slightly longer latencies due to desensitization, consistent with the modest attenuation of feedback sensitivity observed here among habitual users.

H2: Subverse-response. Contrary to H2, the interaction between lagged net votes and subverse type ($\hat{\beta}_{S:SVtype} = 0$, 95% CI [-0.01, 0.01], $p = 0.57$) and the three-way interaction with habit ($\hat{\beta}_{S:H:SVtype} \approx 0$, $p > 0.1$) were not significant. This indicates that community alignment (CT vs non-CT) does not meaningfully modulate reward sensitivity among habitual users. Minor trends suggested bursty users in non-CT subverses posted slightly faster ($\hat{\beta}_{S:H:SVtype} = -0.01$, 95% CI [-0.03, 0.00], $p = 0.07$).

In sum, our results demonstrate that habitual participation systematically reduces the responsiveness of posting behavior to prior social feedback, consistent with predictions from reinforcement learning and habit theory: **users with stronger posting routines increasingly act on internalized cues rather than immediate rewards**. This micro-level attenuation of feedback sensitivity illustrates the transition from reinforcement-driven engagement to stimulus-driven, automatic behavior. Importantly, **subverse alignment with CT norms exerts minimal influence on these dynamics**, suggesting that routine formation, rather than community conformity, dominates short-term posting decisions. These results link temporal patterns of reinforcement sensitivity to habit formation, providing a notion for why highly active users maintain participation even when social rewards fluctuate.

RQ2: Which Factors Predict Upshifts Toward Conspiracy Theory Communities?

Building on short-term responsiveness (RQ1), **RQ2** shifts to the **meso-level**: the weekly process of exploration and immersion into communities. Rather than asking whether users stay active or how fast they return, this question asks *where* they go—specifically, what factors predict immersion into conspiracy theory (CT) subverses.

We integrate three complementary mechanisms. First, social learning and identity frameworks suggest users gravitate toward spaces that better match their beliefs (Bandura and Walters 1977) and where they find similar others (TeBlunthuis et al. 2022).

Second, cognitive–motivational accounts highlight the role of epistemic emotions—such as curiosity and uncertainty—in driving exploratory search. Epistemic emotions orient toward information gain rather than immediate rewards: they motivate inquiry and sustain engagement until uncertainty is resolved (Berlyne 1966). Classic work by Berlyne (1966) proposed that curiosity follows an "optimal arousal" principle, where novelty, complexity, and uncertainty maximize positive valence and exploratory behavior. More recent work shows that exploration is strongest when users face high uncertainty about what to expect, but the signals they encounter are relatively clear (Yanagisawa and Honda 2025). In the context of CT communities, this may translate into broader exploration of narratives and alternative explanations, as users seek to reduce uncertainty while still encountering novel or provocative claims.

Third, habit theory implies that repeated posting routines structure exposure and shape the likelihood of change. These behaviours might be motivated through distinct cues in a

text: some signals open cognitive "loops" that prolong interpretive search, whereas others supply attributional closure that anchors users within established narratives.

We therefore distinguish two classes of narrative signals. *Trait cues* (secrecy, pattern, threat) are message-centric cues that increase perceived uncertainty, hiddenness, or danger. Such cues could trigger epistemic emotions—curiosity, doubt, or unease—that heighten arousal and motivate further search for alternative explanations. This need for mitigating uncertainty is consistent with Uncertainty Reduction Theory (Berger and Calabrese 1974).

By contrast, *group cues* (actor, action) assign responsibilities and reference points. These agent-centric signals supply attributional closure, which might reduce uncertainty and further exploratory search. They might function as anchors on established interpretations within the community. Habit theory further suggests that in volatile informational environments, repeated routines may serve as coping mechanisms that "lock in" behavior over time (Gardner and Lally 2018).

We therefore test the following hypothesis:

- **H3: (CT Upshifts).** The probability of an upshift into CT subverses in week t increases with (a) prior exposure to *trait* cues (Secrecy, Pattern, Threat), (b) low social reinforcement (i.e., net votes), and (c) habitual posting. By contrast, a higher prevalence of *group* cues (Actor, Action) in the prior period will be associated with *reduced* probability of additional upshift, consistent with narrative consolidation.

Measures

Outcome: Weekly Upshift to CT Subverse. The binary outcome y_{it} equals 1 if user i exhibits an increase in mean subverse CT score from week $t-1$ to week t and 0 otherwise, i.e.

$$y_{it} = \mathbf{1}\{\Delta \mathrm{SV}_{it} = \mathrm{SV}_{it} - \mathrm{SV}_{i,t-1} > 0\}$$

where SV_{it} is the user-week mean subverse score. This operationalisation captures growth in CT-oriented posting rather than stability.

Content Features: Trait and Group Signals. We detect five multi-label narrative features at the post level (i.e., submission and comments) using a fine-tuned BERT classifier trained on human annotations (Batzdorfer 2024). For analysis, we form two composite dimensions:

- **Trait cues** (T): Secrecy, Pattern and Threat. These are intrinsic message cues (hidden influence, repeated orchestrated patterns, imminent danger) that may raise epistemic uncertainty and search impetus.
- **Group cues** (G): Actor and Action. These are agent-centered cues that name coalitions or actors and describe coordinated behavior; in our annotations, these specifically capture actors framed as malicious or instrumental in harmful activity.

At the post level, each feature is coded as a binary variable. A post is coded as expressing *group cues* if actor and action are present; a contribution is coded as expressing a *trait* pattern if secrecy, threat, and pattern are detected. These post-level binaries are aggregated to user-week means and then lagged by one week (e.g., $T_{i,t-1}$, $G_{i,t-1}$) to capture prior exposure. For construct validity of the content features, see Appendix A.

Habit and Reinforcement. Posting routine is captured by a three-level *habit class* (occasional, steady-regular, bursty) derived from users' temporal posting patterns. Net votes are included as a lagged predictor.

Model

We modeled weekly transition probabilities using a generalized additive mixed model (GAMM) with a binomial (logit) link. The model included the lagged outcome ($y_{i,t-1}$) to capture short-term inertia; penalized smooth functions for prior trait cues ($T_{i,t-1}$), group cues ($G_{i,t-1}$), and net votes ($S_{i,t-1}$) to allow nonlinear effects; covariates for week index (W_t) and weekly comment count (C_{it}) to adjust for temporal trends and overall activity; habit class (H_i) as a fixed effect; and random intercepts for users (u_i) to account for repeated measures (Eq. 2). GLMM results are provided in Appendix B as robustness checks.

$$\begin{aligned} y_{it} &\sim \mathrm{Bernoulli}(p_{it}), \\ u_i &\sim \mathcal{N}\left(0, \sigma_u^2\right), \\ \mathrm{logit}(p_{it}) &= \alpha \;+\; \gamma\, y_{i,t-1} \;+\; f_1(T_{i,t-1}) \;+\; f_2(G_{i,t-1}) \\ &\quad +\; f_3(S_{i,t-1}) \;+\; f_4(W_t) \;+\; f_5(C_{it}) \\ &\quad +\; \beta_{\mathrm{SR}}\, \mathbb{I}\{H_i = \mathrm{SR}\} \;+\; \beta_{\mathrm{B}}\, \mathbb{I}\{H_i = \mathrm{B}\} \;+\; u_i \end{aligned} \tag{2}$$

Here, $y_{it} = 1$ if user i shows a weekly CT upshift at week t; $y_{i,t-1}$ is the lagged upshift; $T_{i,t-1}, G_{i,t-1}, S_{i,t-1}$ are lagged trait, group, and net vote scores; H_i is habit class (SR = steady regular, B = bursty); W_t is week index; C_{it} is weekly comment count; $f_j(\cdot)$ are smooth functions; u_i is a user-level random intercept.

Results

Random intercepts for users accounted for substantial between-user variation ($edf = 2325.51$, $\chi^2 = 20{,}484$, $p < .001$), confirming strong heterogeneity across individuals.

Parametric Effects. Users' upshift behavior exhibited strong **temporal inertia**: a prior upshift markedly suppressed the likelihood of another upshift in the following week ($\hat{\gamma} = -1.18$, $SE = 0.012$, $z = -102.6$, $p < .001$). Habit class significantly structured transition dynamics: both steady regular ($\hat{\beta}_{SR} = 0.70, p < .001$) and bursty users ($\hat{\beta}_B = 0.68, p < .001$) exhibited a higher baseline probability of CT upshift compared to occasional contributors. This pattern suggests that habitual engagement—whether consistent or surge-like—is associated with **increased susceptibility to upshifts** into CT subverses.

Smooth Effects. Among the content predictors, only group-related discourse showed a significant nonlinear effect (edf $= 7.1$, $\chi^2 = 319.3$, $p < .001$), with higher levels

of group cues corresponding to a reduced likelihood of upshift. Trait-related content cues (edf $= 2.2$, $\chi^2 = 3.7$, $p = 0.25$) and net votes (edf $= 1.0$, $p = 0.76$) were not significant predictors. Temporal (edf $= 7.5$, $p < .001$) and activity (edf $= 4.9$, $p < .001$) smooths captured baseline seasonal and volume effects.

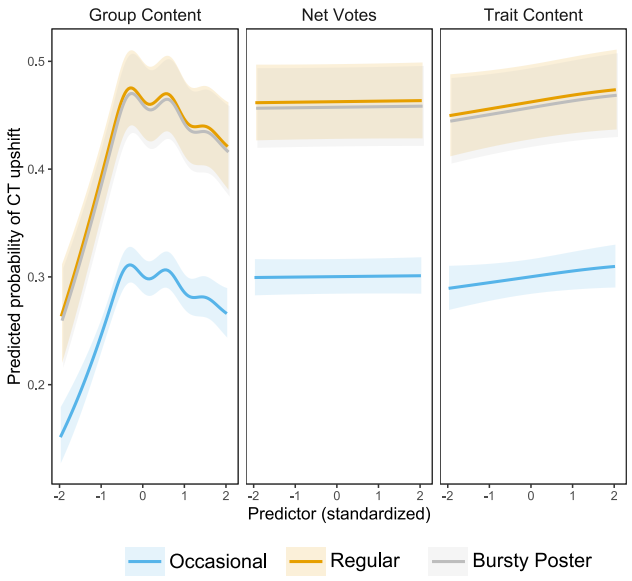

Figure 3: Marginal probability of CT subverse upshift by content cues and net votes, stratified by posting routine. Curves are GAMM-derived population-level predictions for standardized (lagged) trait content, group content, and net votes; shaded bands are 95% Wald intervals computed on the logit scale and transformed to probability. Non-focal covariates fixed at reference values (0 for standardized predictors). User random effects are excluded.

Marginal predicted probabilities (Fig. 3) across standardized content features revealed the probability of transitioning into CT subverses varied between $p \approx 0.15$ and $p \approx 0.48$, depending on habit class and predictor values. Median predicted probabilities were $p \approx 0.39$–0.45. Consistent with the GAMM smooths, only group-related content substantially shaped transitions: higher levels of group attribution were associated with reduced likelihood of CT upshifts (Fig. 3). A linear GLMM (see Appendix A, Fig. 7) with the same covariates (without smooths) yields a consistent qualitative pattern (negative linear effect of G; non-significant T and S), corroborating the GAMM's conclusions. Our findings provide partial support for H3 regarding CT subverse upshifts.

H3a: Trait and Group Cues. Contrary to expectations, exposure to trait cues (secrecy, pattern, threat) did not robustly increase engagement. While trait cues theoretically heighten epistemic arousal and drive exploratory behavior, their effect appears weak in practice, suggesting that uncertainty alone may be insufficient to prompt upshifts without additional motivating factors. Consistent with the stabilizing role of agent-centric signals, higher prior emphasis on group cues (actor, action) **reduced** the likelihood of **CT upshifts**. These cues might provide attributional closure and anchor users within existing communities.

H3b: Social Reinforcement. Net votes exerted minimal influence, indicating that peer feedback does not substantially modulate heightened engagement behavior.

H3c: Habitual Posting. As predicted, steady regular posting and higher activity elevated the baseline probability of upshifts, highlighting the role of habitual engagement in sustaining susceptibility to CT immersion.

The results suggest that week-to-week **migration into CT subverses is structured more by habitual engagement than by exposure to conspiratorial signals per se**. Trait cues (secrecy, pattern, threat) showed no significant association with upshifts, indicating that uncertainty-inducing messages alone do not reliably drive upshifts into CT spaces. In contrast, group cues (actor, action) were negatively associated with upshifts, consistent with theoretical accounts in which agent-centered narratives provide attributional closure and stabilize engagement within existing communities. Habitual posting elevated baseline susceptibility, highlighting the structural role of repeated behavior in shaping community selection. Net voting had a limited association, suggesting that **social rewards may influence engagement intensity but not destination**. Together, these patterns suggest that upshifts are guided less by motivations and epistemic arousal alone but by the interplay of group content framing and behavioral regularities.

RQ3: How Does Persistence Vary with Social Feedback, Routine, and Community?

Finally, **RQ3** moves to the **macro-level**: the persistence of engagement—or its inverse, dropout—over the course of months. Unlike short-term posting latency (RQ1) or shifts into subverses (RQ2), dropout marks a categorical change: users stop participating altogether. This transition is critical because it determines not only individual trajectories but also the survival and cohesion of communities. Studying persistence at this longer timescale allows us to ask questions that short-term analyses cannot: What sustains engagement when immediate rewards are absent? How do routines and community context buffer against participation decline?

From a reinforcement learning (RL) perspective, persistence should be highly sensitive to feedback: positive peer signals (e.g., upvotes) sustain activity, while negative signals (e.g., downvotes, criticism) accelerate disengagement. Yet, persistence in ideologically aligned communities often defies this logic: members continue—and sometimes intensify—their activity even under sparse or hostile feedback. This puzzle suggests that dropout dynamics are governed by mechanisms that go beyond short-term reward responsiveness—namely, habit and ideological identity.

We therefore integrate three perspectives. RL emphasizes how rewards and punishments shape continued activity. Habit theory highlights that routinized behavior can sus-

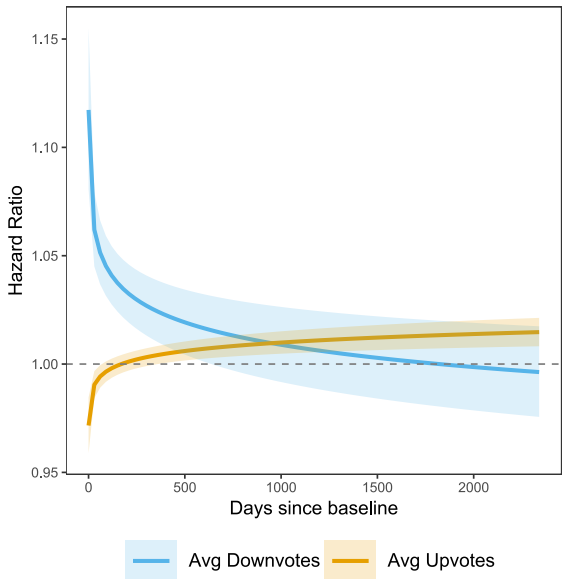

Figure 4: Time-varying effects of social feedback on user dropout hazard. Time-varying hazard ratios for a one-unit increase in average upvotes (purple) and downvotes (green). Shaded areas show 95% confidence intervals; the dashed line at HR = 1 indicates no effect. Downvotes consistently increase dropout risk, while upvotes have a smaller, time-dependent effect.

tain engagement even when reinforcement weakens, as actions become automatic rather than evaluative. Normative alignment with community contexts emphasizes the role of social identity (Tajfel et al. 1971): in cohesive groups such as conspiracy theory (CT) communities, feedback is filtered through ingroup norms, so even negative signals may fail to deter engagement. Together, these perspectives motivate the following hypotheses:

- **H4 (Feedback–persistence):** Users receiving more positive peer feedback (e.g., upvotes) will exhibit a lower hazard of dropout, whereas exposure to negative feedback (e.g., downvotes, controversy) will increase the hazard.
- **H5 (Routine–persistence):** Users with established posting routines (steady or bursty habit) will exhibit lower dropout hazard than occasional posters.
- **H6 (Subverse–persistence):** Users embedded in CT subverses will exhibit a lower dropout hazard compared with users in non-conspiracy subverses.

Measures

We modeled subverse dropout using a Cox proportional hazards regression with time-varying covariates. Proportional hazards were assessed using Schoenfeld residuals, which indicated minor violations for upvotes and downvotes; these were addressed by including explicit time-varying effects in the model.

Outcome: SV Dropout. The outcome was defined as the time to dropout from a subverse, measured in days since a user's first activity in that subverse. To capture evolving engagement patterns while reducing short-term noise, covariates were aggregated over *monthly intervals*: for each user i and subverse s, we computed the mean upvotes, downvotes, controversy, and post count within each month. Each survival interval was defined as a *30-day window* starting from the first day of the month or the user's first comment, whichever applied, truncated at the earliest of 30 days after start, the user's last comment, or 30 days before platform shutdown (2020-12-25) to avoid bias from exogenous closure. An event indicator $e_i = 1$ was assigned if the user ceased commenting within the interval, and $e_i = 0$ otherwise. Upvotes and downvotes were treated as time-varying covariates using a log-transformed time function:

$$g(t) = \log(1 + t)$$

which allows their effect to evolve over time.

Predictors were *posting routines*, *account tenure*, *social feedback*, and *subverse type*.

Social Feedback

Upvotes, Downvotes, Controversiality. We take the ratio of upvotes as a measure of social feedback: letting u and d be the number of net upvotes and downvotes, we calculate the upvote ratio as $\frac{u}{u+d}$. We also consider negative social feedback as the downvote ratio $\frac{d}{u+d}$, as an opposite mechanism signaling the undesirability of users' behavior. Finally, we acknowledge controversial content's ability to elicit polarized reactions in a community, which is expected to reinforce users' behavior positively.

We calculate controversiality as $(u+d)^{\frac{min(u,d)}{max(u,d)}}$ if $u, d > 0$ else 0.[4] This measure is zero if either side of the distribution is absent, but grows with both magnitude and balance, capturing items that elicit simultaneously strong approval and disapproval. Empirically, the distribution is right-skewed: most posts score near zero, while a minority show very high controversy (see Appendix C, Fig. 8). Representative examples are shown in Table 1.

Model

Our Cox proportional hazards regression ($n = 289{,}791$; events $= 14{,}965$), included time-varying covariates for upvotes and downvotes, and a frailty term for subverses to account for unobserved heterogeneity:

[4] As computed on Reddit previously: https://github.com/reddit-archive/reddit/blob/753b17407e9a9dca09558526805922de24133d53/r2/r2/lib/db/_sorts.pyx#L60

$$
\begin{aligned}
b_s &\sim \mathcal{N}(0, \sigma_s^2), \\
g(t) &= \log(1+t), \\
h_{ij}(t) &= h_0(t) \exp\Big(\beta_1 U_{ij} + \beta_2 D_{ij} + \beta_3 U_{ij}\, g(t) \\
&\quad + \beta_4 D_{ij}\, g(t) + \beta_5 H_i + \beta_6 C_{ij} + \beta_7 T_i + \beta_8 G_i \\
&\quad + b_{s[i]}\Big)
\end{aligned}
\tag{3}
$$

Here, $h_{ij}(t)$ is the hazard of dropout for user i in interval j at time t, $h_0(t)$ is the baseline hazard, U_{ij} and D_{ij} are the mean upvotes and downvotes in interval j, H_i is the habit class, C_{ij} is mean controversy in interval j, T_i is tenure stratum, G_i is subverse type (CT vs. non-CT), $b_{s[i]}$ is the subverse-level frailty, σ_s^2 is the variance of subverse frailty, and $g(t)$ transforms time.

Results

The Cox proportional hazards model demonstrated good fit (concordance $= 0.75$), with the likelihood ratio test highly significant ($\chi^2 = 12{,}426$, df $= 1{,}028$, $p < 2 \times 10^{-16}$). The inclusion of subverse-level frailty captured substantial unobserved heterogeneity in dropout rates ($\sigma^2 = 0.80$), indicating that users within some subverses are consistently at higher or lower risk of dropout.

H4: Feedback–persistence. Consistent with the hypothesis, negative feedback increased dropout hazard, whereas positive feedback had a small time-dependent protective effect. Specifically, a one-unit increase in average downvotes raised dropout risk (HR $= 1.12$, 95% CI [1.08, 1.16], $p < 0.001$), whereas upvotes were associated with a modest reduction in risk initially (HR $= 0.97$, 95% CI [0.96, 0.99], $p < 0.001$), with a slight increase over time ($\text{HR}_{tt} = 1.01$, 95% CI [1.00, 1.01], $p < 0.001$) (see Fig. 4). Controversy showed no significant effect (HR $= 1.02$, 95% CI [1.00, 1.04], $p = 0.11$).

H5: Routine–persistence. Habitual posting patterns were associated with higher dropout risk rather than buffering against it. Steady posters had elevated hazard (HR $= 1.15$, 95% CI [1.10, 1.19], $p < 0.001$), and bursty posters even more so (HR $= 1.40$, 95% CI [1.34, 1.46], $p < 0.001$) (see Fig. 9). These findings suggest that high posting frequency, whether steady or bursty, may accelerate disengagement. Thus, routine posting did not confer persistence as hypothesized.

H6: Subverse–persistence. Users participating in conspiracy-themed subverses experienced lower dropout hazard compared to those in non-conspiracy subverses. Being in a non-conspiracy subverse substantially increased dropout risk (HR $= 1.80$, 95% CI [1.57, 2.07], $p < 0.001$) (see Fig. 9), highlighting the potential buffering effect of subverse alignment and suggesting that community context contributes to user persistence.

Taken together, these results suggest a tension between individual habit formation, social feedback, and community context in shaping persistence. First, **negative signals (downvotes) act as localized punishments that increase the hazard of leaving a subverse**, consistent with reinforcement learning. Second, rather than buffering against disengagement, habitual posting suggests accelerating exit. Third, ideological context exerts a countervailing force: **conspiracy-themed subverses anchor users by providing community alignment, reducing dropout hazard even in the face of negative feedback or intense posting**. Overall, these findings point to a model of **conditional persistence**: social punishment pushes users away, habitual activity can destabilize engagement, but community alignment can sustain it. Dropout thus reflects not only disengagement but also the interaction of habits and social signals within the community context.

Discussion and Limitations

Taken together, our analyses shed light on how users turn to conspiracy communities.

Mechanisms of Engagement with Conspiracy Theory Communities

Research on the role of CT users in mainstream social media portrays them as a vocal minority, showing a tendency toward engaging in conflict, toxicity, and overall hyperactivity compared to mainstream users. These behaviors are often interpreted as the end product of a radicalization process, conditioned by an extended engagement with fringe ideas (Faddoul, Chaslot, and Farid 2020). By modeling the transitions of individuals out of mainstream communities and into CT communities, we find that these behaviors may precede rather than result from engagement with CT communities. Individuals who eventually join CT spaces tend to self-select into them based on pre-existing tendencies toward high activity and consistent participation.

This lends nuance to our current understanding of the mechanisms that determine engagement with CT communities. Past research identified how social factors in users' transitions into CT communities may act as both positive and negative forces: positive social interactions, such as engaging with CT members in mainstream spaces, may pull users into CT communities, as well as negative social interactions, such as receiving negative peer feedback and being sanctioned through moderation, may push users out of mainstream spaces (Habib, Srinivasan, and Nithyanand 2022; Phadke, Samory, and Mitra 2021b,a). While we find corroborating evidence for these social forces, by broadening our frame of analysis, we also discover that they are not specific to CT communities, but rather apply to the dynamics of community transitions in general. In other words, social feedback seems less important for upshifts into CT spaces.

The lack of direct effect of social feedback on transitions toward CT communities does not discount, however, their specific significance to CT engagement. Social interactions, both positive and negative, continue to influence engagement, yet their effects appear indirect rather than deterministic. Positive feedback fosters habitual posting, which in turn facilitates the likelihood of engaging with CT communities, whereas negative feedback primarily shapes persistence and

resilience. Habitual engagement emerges as a central driver of CT community participation, aligning with prior findings in analogous contexts, such as the "Manosphere," where ingroup acceptance reinforces behavioral shifts and participation intensity (Habib, Srinivasan, and Nithyanand 2022).

Integrating habit as a persistent factor in users' engagement and going beyond the temporary effect of social feedback provides important insights into the inner workings of CT communities. There, users show significantly higher retention than in mainstream communities: not only do CT users appear to derive stronger rewards from positive social feedback, which relates to high retention, but retention remains high even in the face of negative feedback, controversy, and sanctions, such as when posting privileges are negated. This view mirrors related research on CT communities, which finds them to be resilient to social and technological exclusion, as well as research on individuals, which highlights the self-sealing properties of CT narratives and beliefs against criticism (Monti et al. 2023). Our findings suggest that habitual routines are a key driver connecting short-term rewards to long-term immersion and persistence in CT communities. This raises the question of whether these habits keep users engaged beyond the immediate feedback they receive, and how to intervene.

Interventions on Habit Disruption

Classic learning theory and dual-process accounts from cognitive psychology set important boundaries on what interventions can accomplish. According to the "Law of Effect", behaviors followed by reinforcing consequences strengthen stimulus–response (S–R) associations (Thorndike 1927). Extinction procedures do not erase such associations but create new context-specific learning that competes with the original S–R link; consequently, simply withholding reinforcement (e.g., removing upvotes) often yields only transient reductions in the target behavior and is prone to spontaneous recovery and renewal when contextual cues persist (Buabang et al. 2025). Similarly, the dual-system perspective emphasizes that habits are more likely to be expressed under short response preparation times and when goal-directed control is strained (low attention, cognitive load) (Buabang et al. 2025).

These constraints imply two things for platforms. First, interventions that rely solely on reinforcement (e.g., suppressing or delaying the visibility of upvotes) are likely to be effective only in the early stages of engagement, before routines consolidate. Second, lasting change requires either creating new S-R associations that compete with existing routines or strengthening goal-directed control at moments when users can deliberate.

One way to achieve this could be to adjust the timing and visibility of social feedback during a user's early community participation. For example, delaying or aggregating upvote counts on initial posts may slow the formation of stimulus–response habits and lower the likelihood that users develop steady or bursty posting routines. Another point to intervene is to disrupt the contextual cues that precede posting, such as notifications or time-of-day regularities. Changing these cues makes them less reliable triggers, which in turn dampens the habitual responses they normally elicit. Another strategy could be substitution: encouraging new routines that build directly on old cues (Buabang et al. 2025). If a user opens the platform at the same time each day, the platform could introduce a prompt before posting, thereby stacking a new behavior on an existing routine. Finally, introducing a short delay at the point of posting, such as a confirmation, could expand the brief window in which goal-directed control can act. By giving users more time to prepare their response (Buabang et al. 2025), these nudges may slow posting, reduce rapid within-session reposting, and shift behavior from automatic to more deliberate action.

Limitations. Our work has several limitations. First, the vector-based embedding of subverses along the CT dimension is time-agnostic, aggregating data across the entire user history, which assumes that sub-communities on Voat remain static in terms of CT orientation. Additionally, a significant proportion of Voat users are anonymous and may operate multiple accounts, potentially distorting the accuracy of subverse embedding scores. Moreover, our analysis is limited to behavioral forms of social feedback—while other reward types, potentially inferred from post content (e.g., argument-level), remain unexplored and warrant future research. Additionally, by focusing solely on individual user trajectories, we overlook the interactions within discussion threads, which could provide a more granular perspective on reward mechanisms and offer insights into how subverse dynamics evolve from the thread level to the broader subcommunity. Furthermore, the observational nature of our study limits our ability to draw causal inferences; specifically, regarding habit, it cannot be determined whether behavior is driven by intrinsic motivation, social reinforcement, or pure habit. Furthermore, our operationalization of constructs like controversy or habit relies on platform activity metrics, which may not fully capture underlying psychological mechanisms or be distorted as we lack behavioral measures of lurkers and viewing without behavior.

While our results reveal robust mechanisms linking social feedback, habit, and ideological context to persistence, they should be interpreted with caution. Voat was a niche platform with a distinctive user base and had already shut down at the time of analysis. These factors limit the direct generalizability of our findings to broader populations. At the same time, Voat shared key affordances with many other platforms — voting, threaded discussions, community moderation — making it a useful "critical case" for observing how these mechanisms unfold under extreme conditions of ideological homogeneity. In this sense, our findings are less about Voat per se than about how platform affordances and user routines interact to shape persistence. Future work is needed to test whether the same dynamics appear in mainstream environments, where ideological boundaries are more porous and community contexts are more heterogeneous.

Conclusion

Online conspiracy communities endure not because rewards sustain them, but because habits do. Persistence emerges from the interplay of reinforcement, habit, and

community alignment—and these forces operate at different timescales. Occasional users remain sensitive to feedback, but once routines crystallize, regular contributors continue posting largely independently of external rewards. In conspiracy-themed communities, ideological cohesion further buffers against negative evaluation, fostering resilience even in adverse environments. Positive feedback can redirect less embedded users toward mainstream participation, yet among habitual contributors, it instead entrenches existing routines—a dynamic that fundamentally limits the reach of reward-based platform interventions and calls for habit-disruption strategies instead.

Acknowledgments

This research was partly supported by the Social Media for Democracy (SoMe4Dem) under the European Union under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01-07). Any opinions, findings, and conclusions or recommendations expressed do not necessarily reflect the views of the funding sources.

References

Anderson, I. A.; and Wood, W. 2023. Social motivations' limited influence on habitual behavior: Tests from social media engagement. *Motivation Science*, 9(2): 107.

Bandura, A.; and Walters, R. H. 1977. *Social learning theory*, volume 1. Prentice hall Englewood Cliffs, NJ.

Banisch, S.; Jacob, D.; Willaert, T.; and Olbrich, E. 2025. The social dilemma of online segregation: A dynamical model of platform choice. *Rationality and Society*, 10434631261428850.

Bates, D.; Maechler, M.; Bolker, B.; and Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1): 1–48.

Batzdorfer, V. 2024. Conspiracy Narratives on Voat: A Longitudinal Analysis of Cognitive Activation and Evolutionary Psychology Features. In *Proceedings of the 16th ACM Web Science Conference*, 42–47.

Berger, C. R.; and Calabrese, R. J. 1974. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human communication research*, 1(2): 99–112.

Berlyne, D. E. 1966. Curiosity and Exploration: Animals spend much of their time seeking stimuli whose significance raises problems for psychology. *Science*, 153(3731): 25–33.

Bermúdez, J. P.; and Felletti, F. 2021. Introduction: Habitual Action, Automaticity, and Control. *Topoi*, 40(3): 587–595.

Brady, W. J.; McLoughlin, K.; Doan, T. N.; and Crockett, M. J. 2021. How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33): eabe5641.

Bruckman, A. 2002. Ethical guidelines for research online. https://faculty.cc.gatech.edu/~asb/ethics/. Accessed: 01-03-2026.

Buabang, E. K.; Donegan, K. R.; Rafei, P.; and Gillan, C. M. 2025. Leveraging cognitive neuroscience for making and breaking real-world habits. *Trends in Cognitive Sciences*, 29(1): 41–59.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2014. How community feedback shapes user behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 41–50.

Diab, A.; Nefriana, R.; and Lin, Y.-R. 2024. Classifying Conspiratorial Narratives at Scale: False Alarms and Erroneous Connections. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 340–353.

Douglas, K. M.; and Sutton, R. M. 2022. What Are Conspiracy Theories? A Definitional Approach to Their Correlates, Consequences, and Communication. *Annual Review of Psychology*, 74(1): 271–298.

Faddoul, M.; Chaslot, G.; and Farid, H. 2020. A Longitudinal Analysis of YouTube's Promotion of Conspiracy Videos. 1–8.

Fiesler, C.; and Proferes, N. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society*, 4(1): 2056305118763366.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gardner, B.; and Lally, P. 2018. Modelling habit formation and its determinants. *The psychology of habit: Theory, mechanisms, change, and contexts*, 207–229.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Grinberg, N.; Dow, P. A.; Adamic, L. A.; and Naaman, M. 2016. Changes in engagement before and after posting to Facebook. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 564–574.

Habib, H.; Srinivasan, P.; and Nithyanand, R. 2022. Making a radical misogynist: How online social engagement with the manosphere influences traits of radicalization. *Proceedings of the ACM on human-computer interaction*, 6(CSCW2): 1–28.

Herrnstein, R. J. 1970. On the law of effect 1. *Journal of the experimental analysis of behavior*, 13(2): 243–266.

Korenčić, D.; Chulvi, B.; Casals, X. B.; Toselli, A.; Taulé, M.; and Rosso, P. 2024. What distinguishes conspiracy from critical narratives? A computational analysis of oppositional discourse. *Expert Systems*, e13671.

Levy, O.; and Goldberg, Y. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 302–308.

Lindström, B.; Bellander, M.; Schultner, D. T.; Chang, A.; Tobler, P. N.; and Amodio, D. M. 2021. A computational reward learning account of social media engagement. *Nature communications*, 12(1): 1311.

Mekacher, A.; and Papasavva, A. 2022. " I Can't Keep It Up." A Dataset from the Defunct Voat. co News Aggregator. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1302–1311.

Monti, C.; Cinelli, M.; Valensise, C.; Quattrociocchi, W.; and Starnini, M. 2023. Online conspiracy communities are more resilient to deplatforming. *PNAS Nexus*, 2(10): pgad324.

Papakyriakopoulos, O.; Serrano, J. C. M.; and Hegelich, S. 2020. The spread of COVID-19 conspiracy theories on social media and the effect of content moderation. *The Harvard Kennedy School (HKS) Misinformation Review*, 18.

Phadke, S.; Samory, M.; and Mitra, T. 2021a. Characterizing social imaginaries and self-disclosures of dissonance in online conspiracy discussion communities. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–35.

Phadke, S.; Samory, M.; and Mitra, T. 2021b. What makes people join conspiracy communities? role of social factors in conspiracy engagement. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3): 1–30.

Phadke, S.; Samory, M.; and Mitra, T. 2022. Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In *Proceedings of the international AAAI conference on web and social media*, volume 16, 770–781.

Proferes, N.; Jones, N.; Gilbert, S.; Fiesler, C.; and Zimmer, M. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2): 20563051211019004.

Russo, G.; Ribeiro, M. H.; and West, R. 2024. Stranger Danger! Cross-Community Interactions with Fringe Users Increase the Growth of Fringe Communities on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 1342–1353.

Samory, M.; and Mitra, T. 2018. 'The Government Spies Using Our Webcams' The Language of Conspiracy Theories in Online Discussions. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–24.

Tajfel, H.; Billig, M. G.; Bundy, R. P.; and Flament, C. 1971. Social categorization and intergroup behaviour. *European journal of social psychology*, 1(2): 149–178.

TeBlunthuis, N.; Kiene, C.; Brown, I.; Levi, L.; McGinnis, N.; and Hill, B. M. 2022. No community can do everything: why people participate in similar online communities. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–25.

Therneau, T. M. 2023. The Survival Package. *Journal of Statistical Software*, 105(7): 1–29.

Thorndike, E. L. 1927. The law of effect. *The American journal of psychology*, 39(1/4): 212–222.

Turner, G.; Ferguson, A. M.; Katiyar, T.; Palminteri, S.; and Orben, A. 2025. Old strategies, new environments: Reinforcement learning on social media. *Biological psychiatry*, 97(10): 989–1001.

Van Prooijen, J.-W.; and Jostmann, N. B. 2013. Belief in conspiracy theories: The influence of uncertainty and perceived morality. *European journal of social psychology*, 43(1): 109–115.

Van Prooijen, J.-W.; and Van Vugt, M. 2018. Conspiracy theories: Evolved functions and psychological mechanisms. *Perspectives on psychological science*, 13(6): 770–788.

Waller, I.; and Anderson, A. 2021. Quantifying social organization and political polarization in online platforms. *Nature*, 600(7888): 264–268.

Whitson, J. A.; and Galinsky, A. D. 2008. Lacking control increases illusory pattern perception. *science*, 322(5898): 115–117.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L. D.; François, R.; Grolemund, G.; Hayes, A.; Henry, L.; Hester, J.; Kuhn, M.; Pedersen, T. L.; Miller, E.; Bache, S. M.; Müller, K.; Ooms, J.; Robinson, D.; Seidel, D. P.; Spinu, V.; Takahashi, K.; Vaughan, D.; Wilke, C.; Woo, K.; and Yutani, H. 2019. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43): 1686.

Wood, S. N. 2017. Generalized Additive Models: An Introduction with R. *Chapman and Hall/CRC*.

Yanagisawa, H.; and Honda, S. 2025. Modeling the arousal potential of epistemic emotions using Bayesian information gain: a framework for inquiry cycles driven by free energy fluctuations. *Frontiers in Psychology*, 16: 1438080.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, see the contributions list in the introduction.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see the sample section.
   (e) Did you describe the limitations of your work? Yes, see the limitations section.
   (f) Did you discuss any potential negative societal impacts of your work? No
   (g) Did you discuss any potential misuse of your work? No
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, see introduction.
   (b) Have you provided justifications for all theoretical results? Yes, see background.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes, see research questions.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, see limitation section.

(e) Did you address potential biases or limitations in your theoretical framework? Yes, see the limitations.

(f) Have you related your theoretical results to the existing literature in social science? Yes, see related work.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, see implications.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes

(c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? NA

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Appendices

All analyses were conducted in R (v4.3.2) using the packages `tidyverse` (v2.0.0) (Wickham et al. 2019), `lme4` (v1.1-35.1) (Bates et al. 2015), `survminer` (v0.5.0), `survival` (v3.6-4) (Therneau 2023), `mgcv` (v1.9-0) (Wood 2017), `irr` (v0.84.1), `lubridate` (v1.9.3), `caret` (v6.0-94), and `pROC` (v1.19.0.1).

A Validation of Subverses and Content

Validation of Subverse Scoring. We evaluated subverse scores using a stratified sample of $N = 143$ posts drawn from 60 subverses, comprising the top and bottom 30 communities ranked by mean subverse score. Two annotators coded, and diverging samples were discussed until agreement was reached. Inter-rater reliability was high (Cohen's $\kappa = 0.845$, $z = 10.1$, $p < 0.001$). Subverse score, denoted sv, strongly predicted human-assigned conspiracy-theoretic (CT) labels under both consensus schemes: liberal (OR $= 1.73$, 95% CI 1.48–2.10) and strict (OR $= 1.83$, 95% CI 1.53–2.26). Discrimination was robust, with area under the curve values of 0.79 (liberal) and 0.81 (strict) (Fig. 5c,d). Stricter consensus produced marginally higher predictive validity and a better model fit ($AIC = 140$ vs. 147.5).

Validation of Content Features. Principal component analysis of the five content features indicated a two-component solution explaining 67% of variance (RC1: 40%, RC2: 27%). The first component (RC1, "Conspiratorial Content") loaded strongly on *Threat* (0.86), *Secrecy* (0.73), and *Pattern* (0.76), capturing hiddenness, threat, and recurring patterns in conspiracy-related comments. The second component (RC2, "Agency/Action") loaded on *Actor* (0.84) and *Action* (0.70), capturing who is performing actions and the nature of those actions. Model fit was acceptable (RMSR = 0.15). Building on this structure, we evaluated automated detection of each feature against human labels. The content features were evaluated on a stratified sample of $N = 250$ posts from the full dataset. For each feature $f \in \{\text{Action, Actor, Threat, Secrecy, Pattern}\}$, posts were stratified by subverse, and up to $n = 3$ posts per subverse were sampled without replacement, retaining unique posts across features. Two annotators independently labeled each post-feature pair ($human\_label \in \{0, 1\}$). Consensus

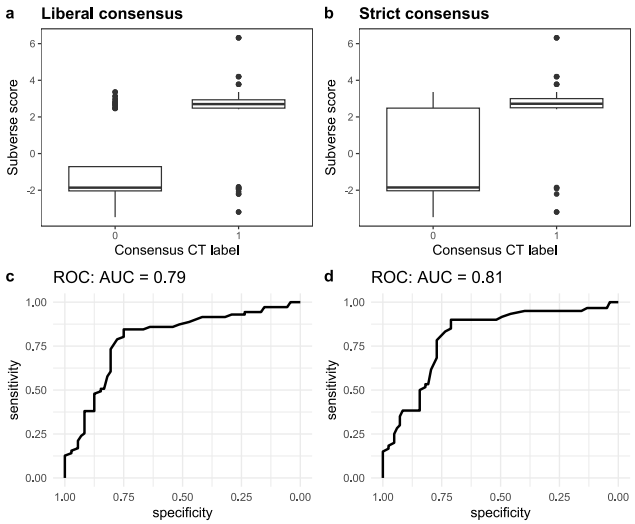

Figure 5: Validation of the subverse score against human annotations of CT content ($n = 143$ posts). (a–b) Distribution of subverse score by consensus labels, showing that higher scores are associated with a greater likelihood of annotators identifying CT content. (a) Liberal consensus (≥ 1 annotator marks CT). (b) Strict consensus (both annotators mark CT). (c–d) Receiver operating characteristic (ROC) curves for the same models, showing robust discrimination (AUC = 0.79, liberal; AUC = 0.81, strict).

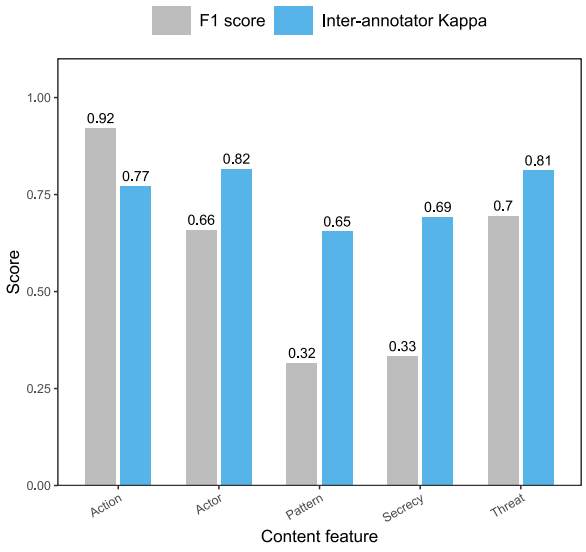

Figure 6: Validation of automated content feature detection against human annotations. Bars show inter-annotator reliability (Cohen's κ) and F1 score for five content features across 250 posts. Human labels were combined using a liberal consensus ((≥ 1 annotator marks CT). Action and Actor features were most reliably detected, whereas Pattern and Secrecy were less consistent.

labels were defined liberally as

$$consensus_f = \begin{cases} 1, & \text{if any annotator marked } f \text{ as present,} \\ 0, & \text{otherwise.} \end{cases}$$

Across 250 posts, automated detection achieved the highest performance for *Action* ($\kappa = 0.772$, Precision $= 0.97$, Recall $= 0.88$, F1 $= 0.92$) and *Actor* ($\kappa = 0.817$, Precision $= 0.50$, Recall $= 0.97$, F1 $= 0.66$) (Fig. 6). Lower performance was observed for *Pattern* ($\kappa = 0.655$, F1 $= 0.32$) and *Secrecy* ($\kappa = 0.692$, F1 $= 0.33$), whereas *Threat* exhibited intermediate reliability ($\kappa = 0.812$, F1 $= 0.70$) (Fig. 6). These results indicate that automated features capture the majority of variance in human-labeled content, with performance varying by feature type.

B GLMM

For reference, we additionally fit a generalized linear mixed-effects model (GLMM) with identical covariates and a random intercept for user, optimized with the bobyqa routine. As shown in (Fig. 7), fixed-effect estimates from the GLMM were substantively consistent with the GAMM results reported in the main text.

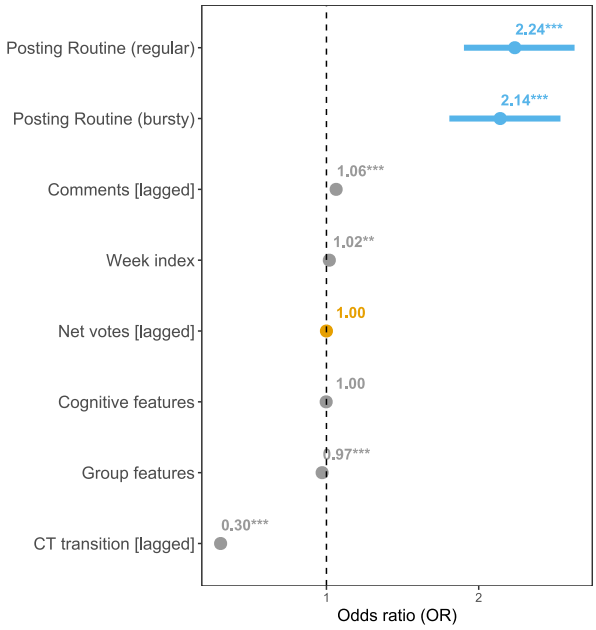

Figure 7: Generalized linear mixed model (GLMM) predicting conspiracy theory activity upshift. Fixed effects estimates shown. Odds ratios with values > 1 indicate an increase of 1 standard deviation. All estimates are adjusted for the autoregressive effect of the prior week's transition status and include random intercepts for users. (Reference level for posting routine is occasional posting.) Point estimates with 95% confidence intervals. Significance levels are denoted as follows: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

Table 1: Representative tokenized examples of low- and high-controversiality posts.

Controv. group	Text snippet	Upvotes	Downvotes	Controv.
Low controversy	"<LOCATION> is a country now. All imported to vote <PARTY>. It's like a <COUNTRY_TYPE> now."	1	0	0
Low controversy	"If this is NOT a troll; <OPTICS>. This post makes you seem like a <PERSON> who thinks <CONSPIRACY> is the reason <EVENT> happened."	0	1	0
Low controversy	"They know they have been caught. That's why they are squirming and resisting. The tides are turning <GROUP>, <MOVEMENT> is upon us. I'm so excited."	10	0	0
High controversy	"Plow through the <NEGATIVE_LABEL> posts. This place will take off momentarily. Also, if your <NEGATIVE_LABEL> posts are deleted, make sure you link them in your next post. Nobody has time to wade through <NEGATIVE_LABEL> like the mods are <NEGATIVE_LABEL> here without proof of a quality post being deleted. 3 2 1.. BLAST OFF!!!!!!"	41	40	73
High controversy	"Your <RELATION> is a <NEGATIVE_LABEL>, and you're both <NEGATIVE_LABEL> for this '<TOPIC>'."	31	31	62
High controversy	"I notice a large amount of <DISINFO> lately about <PERSON> about to <ACTION> on the <GROUP>. <PERSON> is connected to the <GROUP> apparatus in nearly every conceivable way."	30	29	52

C Controversiality

For descriptives on the controversiality measure, refer to (Fig. 8) as well as representative samples (Tab. 1). The histogram shows that most posts have low controversy scores, with a long tail of highly controversial posts, highlighting the skewed distribution of engagement polarity.

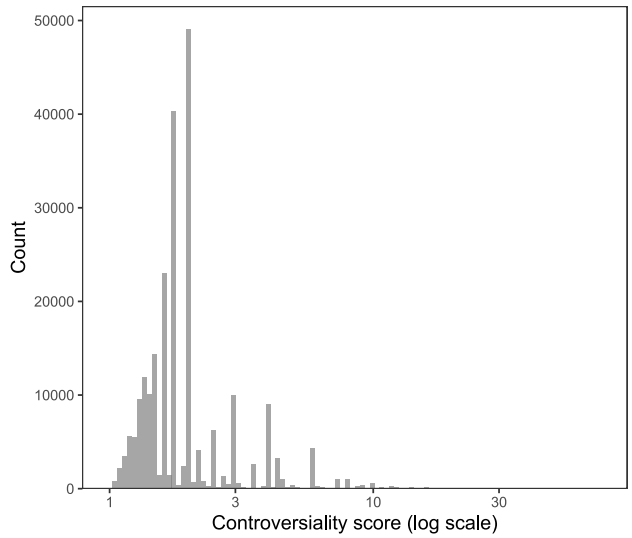

Figure 8: Distribution of the controversiality score for posts with non-zero upvotes and downvotes. The score combines the magnitude and balance of positive and negative feedback.

D Cox Proportional Hazard Model

Hazard Ratios of the Cox proportional hazard regression model (see Fig. 9).

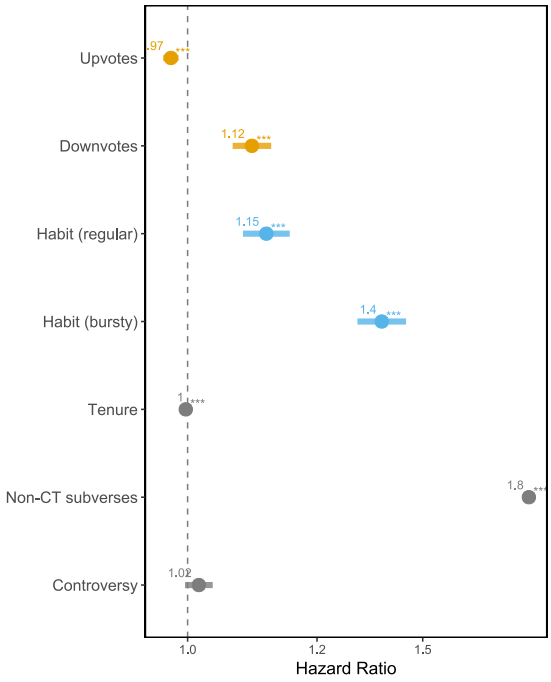

Figure 9: Hazard ratios (HR) for user dropout from the time-varying Cox model. Points indicate HR estimates for the main predictors; horizontal bars represent 95% confidence intervals. Colors indicate predictor category. Stars denote significance ($p < 0.05$; $p < 0.01$; $p < 0.001$). HR > 1 indicates higher dropout risk.